Received April 4, 2021, accepted May 4, 2021, date of publication May 14, 2021, date of current version May 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3080617

Reinforcing Synthetic Data for Meticulous Survival Prediction of Patients Suffering From Left Ventricular Systolic Dysfunction

MOHAMMAD FARHAN KHAN[1], RAJESH KUMAR GAZARA[2,3], MUAFFAQ M. NOFAL[4], SOHOM CHAKRABARTY[2], ELHAM M. A. DANNOUN[5], RAMI AL-HMOUZ[6], AND M. MURSALEEN[7]

[1]School of Water, Energy and Environment, Cranfield University, Cranfield MK43 0AL, U.K.
[2]Department of Electrical Engineering, Indian Institute of Technology Roorkee, Roorkee 247667, India
[3]Centre for Brain Research, Indian Institute of Science, Bengaluru 560012, India
[4]Department of Mathematics and General Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia
[5]General Science Department, Woman Campus, Prince Sultan University, Riyadh 11586, Saudi Arabia
[6]Department of Electrical and Computer Engineering, Sultan Qaboos University, Muscat 123, Oman
[7]Department of Medical Research, China Medical University Hospital, China Medical University, Taichung 40402, Taiwan

Corresponding author: M. Mursaleen (mursaleenm@gmail.com)

ABSTRACT Congestive heart failure is among leading genesis of concern that requires an immediate medical attention. Among various cardiac disorders, left ventricular systolic dysfunction is one of the well known cardiovascular disease which causes sudden congestive heart failure. The irregular functioning of a heart can be diagnosed through some of the clinical attributes, such as ejection fraction, serum creatinine etcetera. However, due to availability of a limited data related to the death events of patients suffering from left ventricular systolic dysfunction, a critical level of thresholds of clinical attributes cannot be estimated with higher precision. Hence, this paper proposes a novel pseudo reinforcement learning algorithm which overcomes a problem of majority class skewness in a limited dataset by appending a synthetic dataset across minority data space. The proposed pseudo agent in the algorithm continuously senses the state of the dataset (pseudo environment) and takes an appropriate action to populate the dataset resulting into higher reward. In addition, the paper also investigates the role of statistically significant clinical attributes such as age, ejection fraction, serum creatinine etc., which tends to efficiently predict the association of death events of the patients suffering from left ventricular systolic dysfunction.

INDEX TERMS Pseudo reinforcement learning, $k-$nearest neighbours, heart failure, synthetic data, support vector machine.

I. INTRODUCTION

Cardiovascular disease (CVD) is one of the proliferated abnormal condition that negatively affects the structure as well as function of the heart and blood vessels. Some of the diseases, such as heart failure, stroke, congenital heart defects, abnormal heart rhythms and peripheral artery disease lies in the larger subset of CVDs. According to the World Health Organization (WHO), CVDs are actively contributing towards global mortality and morbidity. In coming decade i.e. from 2020 to 2030, the death share due to CVDs is likely to pile-up from 31.5% to 32.5% [1], resulting into additional 3.7 million deaths worldwide. The 1%, increment in the overall death share is due to several factors that elevates the risk of developing CVDs, such as age, diabetes, anaemia, smoking, ejection fraction and serum creatinine etc.

In general sense, the co-existing medical conditions (comorbid) such as diabetes, anemia and high blood pressure along with aging are some of the major risk factors for CVDs [2], [3]. The studies have reported that the aging process causes the progressive loss of physiological development at molecular, cellular and tissue levels which leads to an increase in CVDs [4]. On the other hand, several seminal researchers explored the link between CVDs and diabetes, which shows that CVDs are most prevalent cause of death in diabetic patients [5]–[7]. The level of the glucose gets high

The associate editor coordinating the review of this manuscript and approving it for publication was Rosalia Maglietta.

in diabetic patients which can damage the artery wall and increases deposition of fat in the coronary arteries, which might lead to possible heart arrest.

Another well known factor which adversely impacts the functioning of a heart due to CVDs is anemia. In anemia condition, a fall in the level of haemoglobin (Hb) results into inefficient oxygen carrying capacity of the blood that eventually increases mortality in CVDs [9], [10]. Similar to a low haemoglobin condition, high hematocrit or Hb levels is also linked to CVDs [11], [12]. In case of hematocrit, body makes too many red blood cells that makes the blood thicker and ultimately lead to clots, heart attacks, and stroke [11], [12]. Some of the recent studies have directly associated the role of anemia in the dilation of left ventricular leading to left ventricular systolic dysfunction (LVSD) [13], [14].

The ejection fraction is one of the well known factors which distinguishes LVSD in the patients by measuring possible anomaly in the percentage of blood leaving the heart whenever it contracts. In UK alone, approximately 120,000 peoples hospitalise each year due to LVSD [15]. Other factors that are also important in analysing the functioning of the heart are serum creatinine, serum sodium, creatine phosphokinase (CPK), smoking behaviour, platelets etc. [16]–[21]. The aforementioned factors are vital indicator in anticipating the abnormality of the heart functioning which can be used to accurately predict the survival of the patients with the help of *in–silico* models [22].

The development of an *in–silico* computational technologies are vital for predicting the death events of patients suffering from LVSD with the help of various health characteristics of the patients. For example, features such as levels of serum creatinine, serum sodium and CPK enzyme in the blood can be used for identification and classification of possible healthy functioning of the heart. However, it is worth noting that the prediction accuracy of *in–silico* technologies solely depends on the singularity of each feature of the patients in the group with respect to other groups of patients.

SVM model is one of the most broadly employed *in–silico* machine learning algorithms that is used for the bi-classification of events. The core idea of SVM algorithm is to estimate the classifying hyperplane and maximize the margin between the events. There are some major obstacles in using machine learning algorithms in bioinformatics; out of which one of the widely known limitation in a complex real-world dataset is to accurately determine a unique feature of a class which can efficiently distinguish it from the remaining classes [23]. Another highly significant limitation is the inbuilt skewness within the training classes, which tends to influence the overall training process of the machine learning algorithms by inflicting higher bias towards majority class [24].

To overcome the above limitations, one the core objectives of this paper is to observe a subset of clinical attributes which are likely to contribute significantly in estimating a probability of the death event. The additional advantage of estimating the significant sub-factors is that it would act as support system for the doctors, which can help in diagnosing the early signs of possible survival event of the patient due to LVSD. Another objective of this paper is to enhance the prediction accuracy of the death events by reducing the skewness in the training dataset with the help of pseudo reinforcement learning.

This reminder of the paper is organized as follows. A brief overview of the dataset and pseudo reinforcement learning are described in Section II. A statistical significance of the feature space and generation of synthetic dataset is discussed in detail in Section III, followed by concluding remarks in Section IV.

II. MATERIALS AND METHODS

A. DATABASE

The dataset is broadly classified into death events of the patients suffering from left ventricular systolic dysfunction (LVSD) and falling in NYHA class III and IV [4]. The dataset comprised of medical records of 299 heart patients lying within the age group of 40-95 years, has been collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan) in the span of 9 months starting from April 2015 to December 2015 [4].

The survival of the patients has been recorded into categorial binary events, i.e. positive instances or no-death event is recoded as binary-zero, while the dead instances of the patients is recorded as negative instance and represented as binary-one. It is worthy to note that, the dataset is skewed towards negative death instances compared to positive instances i.e. in total only 96 cases have been recorded compared to 203 no-death cases, which is one of the major limitations that may lead to Type-II error in machine learning algorithms and hence can result into inaccurate prediction of death instances

To predict the death events, the dataset consists of twelve features, out of which six features namely anaemia, diabetes, high blood pressure, gender and smoking are category features which have been represented in the binary form, while all the remaining features have been recorded in numeric form. To glance the distribution of all the individual features in the feature space have been illustrated in the form of raincloud plots in Supplementary file S1, which is a combination of raw data points, boxplot, and kernel distribution [25], [26].

In order to visualise the role of comorbidity in the patients, Fig. 1 has been plotted by focusing on age and gender of the patients, because aging is one of the well known lemma which suggests significant relation between comorbid and death events. The small size of circles in the scatter plot represents female patients, while large circles represents male patients. On the other hand, the green colour represents no-death events, while blue represents recorded death events.

Fig. 1(a) distinguishes the role of low haemoglobin level and age factor on the death event of the patients, which suggests that 20 females (out of total 34 deceased females) and 26 males (out of total 62 deceased males) lying in a

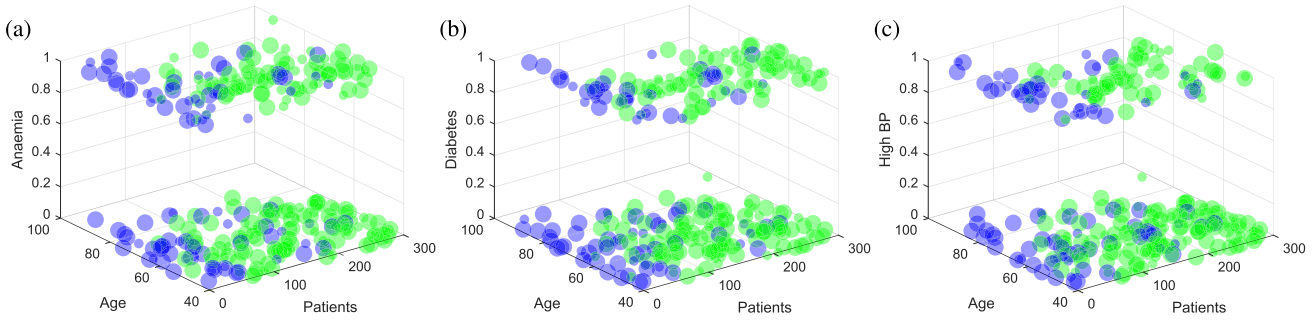


FIGURE 1. Dispersion of patient data suffering from LVSD along with additional medical condition namely: (a) Anaemia, (b) Diabetes, and (c) High blood pressure; with respect to gender and age.

similar age group of 42-95 and 45-95 respectively are unable to survive, and had been suffering from LVSD along with low haemoglobin concentration levels. Similarly from Fig. 1(b), it can be observed that the 20 diabetic female patients of age group 42-82 years, and same number of diabetic male patients from the age group of 45-94 years are unable to survive due to LVSD disorder. In Fig. 1(c), there are 17 deceased female and 22 male patients due to hypertension and LVSD, which are lying in the age group of 46-95 and 45-94 years respectively.

B. PSEUDO REINFORCEMENT LEARNING MODEL

As stated earlier, the core objective of the work reported in this paper is to deal with the problem of data skewness or imbalance in training data which tends to worsen the performance of the machine learning models specifically in terms of prediction of minority class. In order to deal with the limitation of class skewness in the dataset, a reinforcement learning approach has been adopted [27]; which considers the original data space as environment and takes action by continuously populating that data space by appending it with the synthetic data (also called pseudo state). In this work, $k-$nearest neighbour ($k-$NN) algorithm has been used as an agent which takes an action by predicting the possible class of randomly generated synthetic data points, and then appending it in the environment by registering the reward of that specific action.

Analogous to the classical reinforcement learning [28], the proposed agent has a goal to develop a synthetic data by continuously sensing the state of the dataset (environment) and taking the appropriate action to influence the dataset. It is worthy to note that, in the proposed supervised learning problem, the reward process in the advantage table ($\mathcal{A}$-table) is dependent on the coefficient of Goodman and Kruskal's gamma (or γ coefficient) [29] of the statistically significant features. To obtain the specific set of features which significantly define the relation with the death events, three statistical measures namely χ^2 test for independence, Kruskal Wallis one-way ANOVA, and Mann Whitney U test have been adopted [29].

The reward value ($\mathcal{R}$) of pseudo reinforcement learning depends on γ coefficient value and ranges from $+1$ to -1. When the value of γ coefficient for a synthetic dataset is perfectly consistent with the original dataset, then the reward in $\mathcal{A}$-table will be recorded as $+1$, else the reward will proportionally vary to as low as -1 for perfect negative correlation. The termination of the state-action process solely depends on the dilution of the data imbalance problem, that is, the appending process of a minority class would continue through pseudo reinforcement learning until the new synthetic data space has attained a balance between the positive and negative classes.

III. RESULTS AND DISCUSSION

A. AALEN's ADDITIVE REGRESSION MODEL

To further strengthen the lemma, the individual impact of age as well as all the remaining features have been analysed on the survival of the patients through Aalen's additive risk model, which estimates the influence of covariates in impacting the survival of patients over follow-up time [30].

Let the dataset contain information of rightly censored death event ϵ of total n individuals, where each individual is defined via f number of covariates. Let, the complete training dataset $\mathcal{D}$ can be considered as combination of minority and majority classes i.e. $\mathcal{D} := \mathcal{D}_{min} \cup \mathcal{D}_{maj}$; and the information of ith individual in $\mathcal{D}$ can be defined in the triplet form as: $[t_i, \epsilon_i, d_i(t)]$, where $0 \leq t \leq t_i$, t_i is follow-up time, $d_i(t)$ is a feature space of ith patient and $i = 1, 2, \cdots, n$.

For ith individual, the conditional hazard rate h for a given feature $\mathcal{D}_i = (d_{i1}, d_{i2}, \cdots, d_{if})$ at time t can be represented as:

$$h[t|\mathcal{D}_i] = \beta_o(t) + \sum_{m=1}^{f} \beta_m(t) d_{im} \tag{1}$$

where, $\beta_o(t)$ is baseline hazard. The cumulative hazard function ($\hat{h}$) can be obtained by integrating the conditional hazard rate over time t, i.e.

$$\hat{h}[t|\mathcal{D}] = \sum_{m=0}^{f} d_m \int_0^t \beta_m(u) \mathrm{d}u \tag{2}$$

$$= \sum_{m=0}^{f} d_m B_m(t) \tag{3}$$

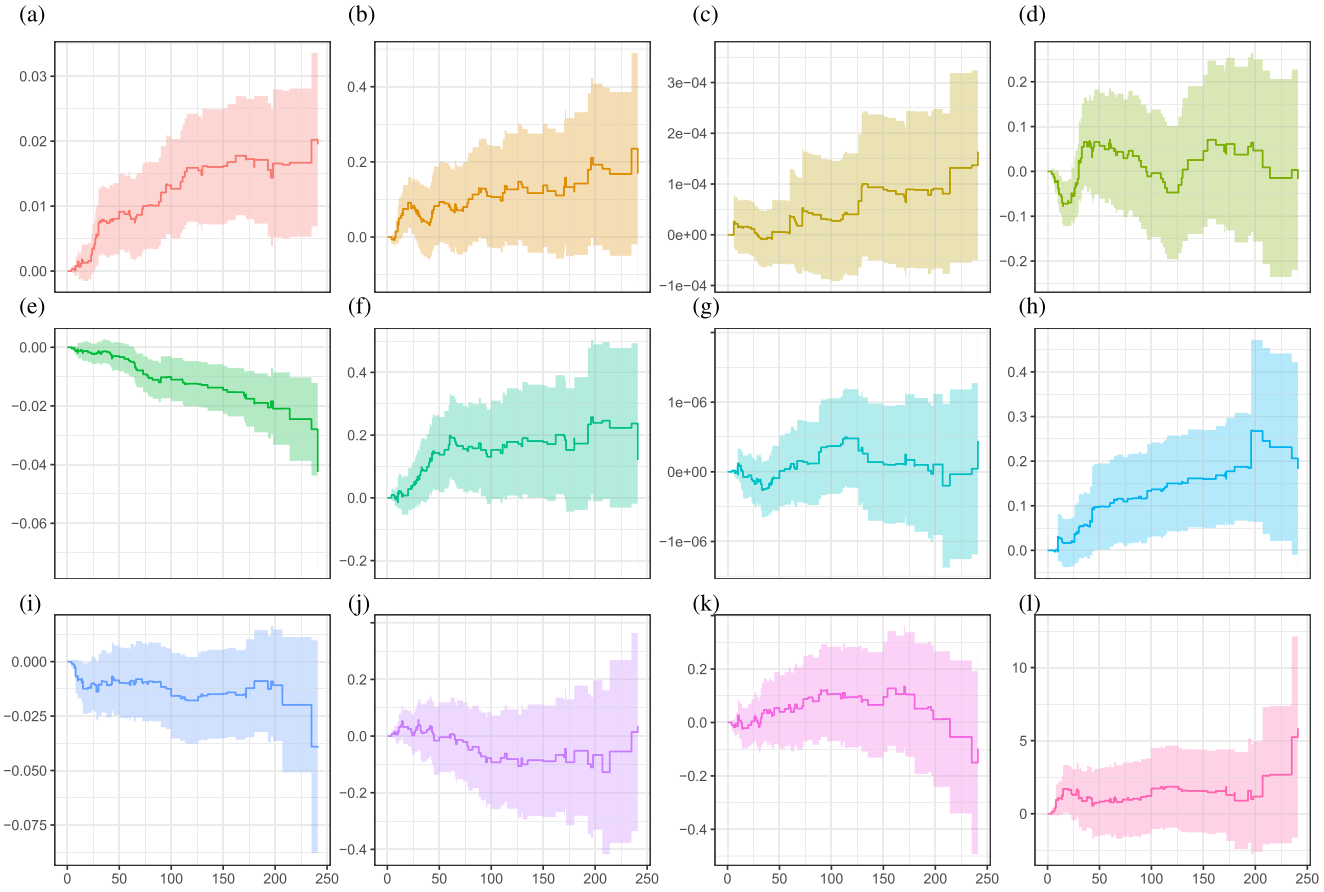


FIGURE 2. Cumulative regression coefficient values of Aalen's additive regression model for: (a) Age, (b) Anaemia, (c) CPK, (d) Diabetes, (e) EF, (f) High BP, (g) Platelets, (h) Serum creatinine, (i) Serum sodium, (j) Gender, (k) Smoking, and (l) Baseline.

where, $d_0 = 1, B_0(t)$ is a baseline cumulative hazard function, and $B_m(t)$ is cumulative regression coefficient for kth covariate [31]. Note that, the additive survival analysis is helpful in extracting the most statistically significant factors whose behaviour should be kept intact while appending the synthetic data.

Fig. 2 illustrates the cumulative regression coefficient values along with 95% confidence limit of Aalen's additive regression model for all the covariates over 241 follow-up days interval. By observing the slope and limits of confidence of cumulative regression coefficient of all the covariates, it can be asserted that only three features namely age, EF, and serum cretinine are playing significant role in predicting hazard associated with the patients. While all the remaining features including baseline subject, the regression curve include origin of cumulative regression coefficient either in lower or upper 95% confidence limit, which suggest no effect on hazard rate due to that specific feature.

Further strengthening the role of aforementioned three factors, Table 1 indicates the Aalen's additive cumulative coefficient of all the features, which further explains the role of the aforementioned three covariate in predicting the hazard level. The exponentiated coefficients of the statistically

TABLE 1. Estimated slope and cumulative regression coefficients ($\hat{h}$) of Aalen's additive regression model.

Covariate	Slope	Coefficient	$p-$value
Age	2.31×10^{-04}	1.81×10^{-04}	0.000331
Anaemia	2.11×10^{-03}	1.77×10^{-03}	0.0619
CPK	1.31×10^{-06}	9.07×10^{-07}	0.274
Diabetes	3.37×10^{-04}	1.53×10^{-04}	0.875
EF	-2.17×10^{-04}	-1.81×10^{-04}	0.00015
High BP	3.22×10^{-03}	2.28×10^{-03}	0.0509
Platelets	2.25×10^{-09}	2.36×10^{-09}	0.628
Serum creatinine	2.71×10^{-03}	2.28×10^{-03}	0.00606
Serum sodium	-2.52×10^{-04}	-2.05×10^{-04}	0.127
Gender	-6.80×10^{-04}	-6.11×10^{-04}	0.589
Smoking	1.24×10^{-03}	5.33×10^{-04}	0.657
Intercept	2.75×10^{-02}	2.49×10^{-02}	0.196

significant covariate in Table 1 can be interpreted as holding a multiplicative effects on the hazard, which are mentioned below:

- Holding the other covariates constant, an additional year of age increases the hazard by a factor of $e^{1.81\times10^{-04}} = 1.0001$ on average, that is, by 0.01%.
- In contrast, higher EF decreases the hazard by a factor of $e^{-1.81\times10^{-04}} = 0.9998$, or 0.02%.

TABLE 2. Results of χ^2 test, Kruskal-Wallis one-way ANOVA and Mann-Whitney U test for estimating impact of all the features on death events.

Features	χ^2 test		Kruskal-Wallis one-way ANOVA			Mann-Whitney U test	
	$p-$value	χ^2 value	$p-$value	K-W χ^2 value	df	$p-$value	W value
Age	0.0044	69.1470	0.0001	14.1783	1	0.0001	7121
Anaemia	0.2728	1.3131	0.2526	1.3087	1	0.2529	9059
CPK	0.3823	209.8449	0.6835	0.1661	1	0.6840	9460
Diabetes	1.0000	0.0011	0.9732	0.0011	1	0.9739	9764
EF	0.0004	65.3315	0.7340×10^{-08}	24.5234	1	0.7368×10^{-08}	13176.5
High BP	0.1809	1.8826	0.1707	1.8763	1	0.1710	8953.5
Platelets	0.6211	172.0794	0.4251	0.6360	1	0.4255	10300.5
Serum creatinine	0.0004	92.4284	0.1573×10^{-11}	40.9352	1	0.1580×10^{-11}	5298
Serum sodium	0.0049	45.8008	0.0002	13.1213	1	0.0002	12261.5
Gender	1.0000	0.0055	0.9406	0.0055	1	0.9412	9787
Smoking	0.8925	0.0476	0.8274	0.0474	1	0.8281	9867
Time	0.0004	245.3263	0.6805×10^{-22}	87.9228	1	0.6852×10^{-22}	16288.5

- Similarly, higher serum creatinine in blood increases the hazard by a factor of $e^{-2.28\times10^{-03}} = 1.0022$, that is, 0.22%.

B. χ^2 TEST, KRUSKAL-WALLIS ONE-WAY ANOVA AND MANN-WHITNEY U TEST

In this section, a role of each feature has been estimated by statistically differentiating the death events of the patients with the help of three statistical tests. Observing Table 2, it can be stated that only five factors namely, age, EF, serum creatinine, serum sodium and follow-up time are significantly defining the relation with death events. Note that, compared to Aalen's hazard model, an additional factor namely serum sodium is also suggested as significant by χ^2 test. In our case, follow-up time cannot be considered as a stand-alone medical feature which has its own existence for providing information on LVSD patients.

To further strengthen the argument about the role of serum sodium in significantly defining the death events, Kruskal-Wallis one-way ANOVA and Mann-Whitney U test have also been considered. Further observing Table 2, it can be asserted that along with serum sodium, the remaining aforementioned factors have rejected the null hypothesis and suggest that the there is a significant difference between the deceased and non-deceased patients which is arising due to variation in the age, EF, serum creatinine, and serum sodium.

C. PSEUDO REINFORCEMENT LEARNING AND SYNTHETIC DATASET

In this step, the limitation of class skewness in the training dataset has been subdued by appending the minority class ($\mathcal{D}_{min}$) of the original datset through pseudo reinforcement learning and transforming it into the synthetic class ($\mathcal{D}^{syn}_{min}$) without largely deflecting the reference γ coefficient value of all the statistically significant features of original dataset. The immediate reward can be estimated as follows:

$$\mathcal{R} = \frac{\sum_{j=1}^{5}\left(1 - \left|\frac{\gamma_j^{obs} - \gamma_j^{ref}}{\gamma_j^{ref}}\right|\right)}{5} \tag{4}$$

where, γ_j^{ref} is reference γ value, and γ_j^{obs} is observed γ value. Ideally the value of $\mathcal{R}$ is unity for the synthetic dataset which is in perfect agrement with the original dataset. The action of the appending process is performed by the $k-$NN agent, which immediately reward the populating process of pre-trained $k-$NN variant. Note that, the agent only picks the majority voted synthetic data point which has been bagged in the favour of minority class within the region of k neighbours, and then append it in the original dataset for estimating the reward of that task.

The three pre-trained variants of $k-$NN i.e. 1, 5, and 15 nearest neighbour(s) are considered in this work. Among considered variants, a simplest one is 1-NN, which predicts the death event $\hat{\epsilon_\vartheta}$ of the unknown individual ϑ (defined via feature $\hat{\mathcal{D}}_\vartheta$) by opting ϵ_i of $\mathcal{D}_i$ lying closest to $\hat{\mathcal{D}}_\vartheta$. The Euclidean distance metric ($\mathcal{E}$) between features ($\mathcal{D}_i, \hat{\mathcal{D}}_\vartheta$) can be estimated as follows:

$$\mathcal{E}_i^2(\mathcal{D}_i, \hat{\mathcal{D}}_\vartheta) = \sum_{m=1}^{f+1}(d_{im} - \hat{d}_{\vartheta m})^2 \tag{5}$$

where, $i = 1, 2, \cdots, n$, and $d_{i(f+1)}$ is follow-up time. The unknown death event $\hat{\epsilon}_\vartheta$ can be predicted as follows:

$$\hat{\epsilon}_\vartheta = \{\epsilon_i | \text{argmin}_{\mathcal{D}_i}\, \mathcal{E}_i^2(\mathcal{D}_i, \hat{\mathcal{D}}_\vartheta)\} \tag{6}$$

Similarly, the other two variants are more generalised $k-$NN network, which estimates the death event associated with $\hat{\mathcal{D}}_\vartheta$ through maximum voting, i.e. the surrounding region comprising k neighbours $N_k(\mathcal{D}_{i'}, \epsilon_{i'})$ that holds maximum probability of occurrence (p) of death event in that region, i.e.

$$\hat{\epsilon}_\vartheta = \{\epsilon_{i'} | \text{argmax}_{\epsilon_{i'}}\, p(\epsilon_{i'} | N_k(\mathcal{D}_{i'}, \epsilon_{i'}))\} \tag{7}$$

where, $(\mathcal{D}_{i'}, \epsilon_{i'}) \in (\mathcal{D}_i, \epsilon_i)$.

After supervising the agents, the pseudo reinforcement learning process generates a set of synthetic point set (ϕ_i^{syn}) by introducing 100 normally distributed data points around each feature $\mathcal{D}_i$. Then the agent utilises a pre-trained knowledge to suggest the class of each synthetic point and bag all the common members to synthetic minority class (ϕ_{min}^{syn}), where $\phi_{min}^{syn} \in \phi_i^{syn}$.

The second step of pseudo reinforcement process initiates by randomly opting an element of ϕ_{min}^{syn} after appending it into $\mathcal{D}_{min}$ and evaluating the reward with respect to perturbation

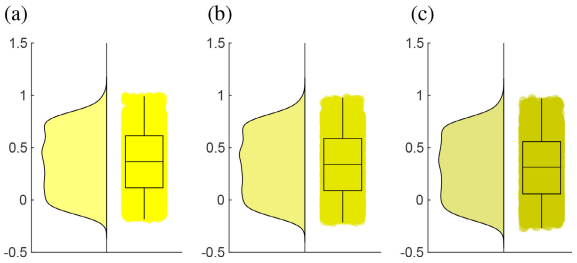


FIGURE 3. Raincloud representation of $\mathcal{A}$-table constituting all the data points introduced around each minority feature $\mathcal{D}_i$ by the agent with: (a) 1 nearest neighbour, (b) 5 nearest neighbours, and (c) 15 nearest neighbours.

in γ^{ref} coefficient value which is defined in eq.(4). The appending action continues until the skewness between the classes reaches close to 90% balance.

It is worth noting that, the reward of the pseudo reinforcement learning process continuously updates itself and solely depends the entire history of the data points appended into the dataset called states. Depending on the relation of reward and action, the agent recommends the variant of $k-$NN to either alter or continue with the priority level of the agent.

Fig 3 represents the distribution of $\mathcal{R}$ within $\mathcal{A}$-table. The columns of the $\mathcal{A}$-table is comprised of three $k-$NN variants, and rows demonstrate the number of synthetic data points populated in the original dataset with respective $\mathcal{R}$ values. It is worthy to note that, according to the box plot and kernel distribution, the upper quartile of all the variants of the agent are constituting 25% of the $\mathcal{R}$ values in the range of [0.6, 1], which suggests the fact that the synthetic data points appended via pseudo reinforcement learning by any of the variant are lying in a close proximity to original dataset and they have largely retained its correlation with original dataset. Further, correlating the performance of all the three agents, it has been found that the pseudo reinforcement learning constituting only a single agent with 1 nearest neighbour is able to achieve slightly better performance, compared to the remaining two agents.

D. PERFORMANCE OF SVM CLASSIFICATION MODEL

In order to evaluate the role of pseudo reinforcement learning in generating the synthetic data which has a tendency to effectively predict the survival rate of LVSD, a variants of SVM classification models has been developed. The SVM model can predict a new information based on its training inputs by differentiating the survival class of the patients through optimal classification boundary [32]. For a given training set (x_i, y_i), $i = 1, 2, \cdots, n$ where $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$, the SVM minimizes the gap ($d = 2/||\vec{w}||$) between two hyperplanes H_1 and H_2 by maximising the modified dual Lagrange objective function, i.e.:

$$\max \sum_{i=1}^{n} \beta_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_i \beta_j y_i y_j (x_i \cdot x_j)$$
$$\text{s.t. } \hat{C} \geq \beta_i \geq 0 \; \forall i, \quad \sum_{i=1}^{n} \beta_i y_i = 0 \tag{8}$$

where, β is Lagrange constraint, and $\hat{C}$ is a soft control variable. In order to deal with the problem of non-linear classification, the expression $(x_i \cdot x_j)$ modifies to $K(x_i, x_j)$, where K represents kernel function. In this work, the evolution in the performance of SVM is computed for four type of kernels, that are linear, quadratic, cubic, and radial basis function (RBF).

Table 3 represents the performance of SVM model trained over skewed original dataset of 299 heart failure patients for classifying the death events. The overall dataset is broadly classified into three forms, namely original, redundant, and synthetic datasets. To curtail the bias of the majority class, the redundant dataset is the one which uses the same copies of the data points exists in the minority class. The other possibility of reducing bias of majority class can be achieved by trimming the random data points of the majority class. However, the trimming process retrenches the bias at the cost of losing the vital information which might be available in the majority class. Hence, the non-feasible trimming process has not been performed in this study.

By utilising entire features of the dataset, the SVM model is able to achieve a best accuracy of 81.27% for the linear kernel (refer Table 3). In contrast, the two of the possible ways to reduce the skewness of the dataset is to either embed the redundant data points of minority class in the dataset or embed the new data points via scientific algorithms like reinforcement learning.

To reduce the skewness of the dataset, the redundant method is likely to use the partially same information for 10-fold cross validation which has already been used in training dataset; hence deeming the redundant method as scientifically flawed. To estimate the impact of such redundancy, a monte-carlo simulation has been performed on a redundant dataset which has been randomly partitioned into 10-folds; and the mean value of the performance metrics is used as final aggregate value for comparison. In Table 3, the best SVM variant trained over a redundant dataset is RBF, which is able to achieve an accuracy of 83.79%. Note that, in Table 3 and Table 4, the top two performing metrics are represented with boldface.

The synthetic dataset developed using pseudo reinforcement learning is giving best accuracy of 87.01% for the linear kernel. It is worthy to note that a quick comparison of the accuracy of all the respective kernels with different datasets reveals that the synthetic dataset has outstandingly outperformed and is giving best accuracy for linear and RBF kernels. A similar trend can be observed in Table 4, in which the RBF kernel is able to train the model in a best possible way. However, the overall glance of Tables 3 and 4 suggest that, opting a subset of statistically significant feature might have reduced the computational complexity of the training process but have worsen the accuracy of the model. Such behaviour is highlighting a fact that, the combination of some of the features which are not statistically significant, are also actively contributing in enhancing the accuracy of the SVM model.

TABLE 3. Performance of four variants of the SVM model trained over entire dataset.

Data type	SVM kernel	Performance metrics				
		Accuracy (%)	AUC	Sensitivity (%)	Specificity (%)	MCC
Original dataset	Linear	81.27	0.8957	80.83	72.72	0.5619
	RBF	78.59	0.9878	81.73	70.00	0.4905
	Cubic	74.58	0.9947	77.09	61.36	0.4047
	Quadratic	73.24	0.9921	78.34	59.75	0.3640
Redundant dataset	Linear	82.74	0.8978	81.14	80.43	0.5946
	RBF	83.79	0.9864	81.77	81.06	0.6780
	Cubic	81.83	**0.9982**	73.14	81.81	0.5450
	Quadratic	81.62	0.9961	76.55	82.16	0.6017
Synthetic dataset	Linear	**87.01**	0.9493	**83.40**	**89.01**	**0.7442**
	RBF	**86.29**	0.9883	**83.79**	**91.19**	**0.7276**
	Cubic	80.28	0.9931	81.67	81.95	0.6063
	Quadratic	80.52	**0.9979**	79.77	82.35	0.6113

TABLE 4. Performance of four variants of the SVM model trained over subset of statistically significant features.

Data type	SVM kernel	Performance metrics				
		Accuracy (%)	AUC	Sensitivity (%)	Specificity (%)	MCC
Original dataset	Linear	74.26	0.8398	**82.43**	74.02	0.5287
	RBF	78.26	0.9465	79.74	73.13	0.4722
	Cubic	81.27	0.8720	81.27	81.25	0.5492
	Quadratic	80.93	0.8818	81.58	76.00	0.5439
Redundant dataset	Linear	73.41	0.8398	75.78	71.21	0.4699
	RBF	77.46	**0.9619**	75.67	85.76	0.5503
	Cubic	75.69	0.8798	71.48	82.87	0.5250
	Quadratic	75.69	0.9157	73.56	78.57	0.5157
Synthetic dataset	Linear	75.96	0.8802	80.68	75.33	0.5190
	RBF	**83.65**	**0.9512**	**82.26**	**89.61**	0.6364
	Cubic	81.73	0.9220	81.97	84.42	**0.6811**
	Quadratic	**82.69**	0.9216	77.87	**88.95**	**0.6627**

Similarly, for entire feature set, the RBF and linear kernels are giving comparatively better results for synthetic dataset for three metrics namely sensitivity, specificity and MCC; while The AUC value of synthetic dataset is not among the top two, but is comparatively equivalent to them. In contrast, for statistically significant feature subset, the RBF kernel is giving comparable performance for synthetic dataset in terms of AUC, sensitivity and specificity.

Investigations have revealed that, generally in bioinformatics, balancing the skewness of the classes prior application of machine learning algorithms is not been a common practise which can dreadfully affect the performance accuracy of the machine learning models by introducing statistical imprecision such as type I and type II errors. A low percentage of specificity and sensitivity for original dataset in Table 3 signify higher type I and type II errors.

Whilst the pseudo reinforcement learning has comparatively reduced both the statistical errors and notably increased the performance of the machine learning models. The RBF kernel of the synthetic dataset gives best performance compared to other SVM kernels which is evident from the highest number of boldfaced values. Similarly in Table 4, the RBF kernel again gives best performance, while deteriorated the overall subjective values due to exemption of some of the features which are significant for a better classification.

It is worthy to note that the importance of statistically significant feature subset can be observed by comparing the rows of original dataset in Tables 3 and 4. It is the case, when appending of synthetic data is not in a priority. Under such scenario, the cubic kernel of the SVM model applied over statistically significant features is able to deliver better accuracy, sensitivity, specificity and MCC. Compared to entire dataset, the statistically significant features are able to give a substantial difference of $+6.69\%$ in accuracy, $+4.14\%$ in sensitivity, $+19.89\%$ in specificity, and 0.1445 extra positive correlation in MCC. Hence, either the machine learning model developed from a subset of significant features can be employed by the doctors, or the subset of features can be used as standalone determinant in efficiently anticipating a possible death instance based on its critical threshold levels.

IV. CONCLUSION

In this paper, a novel pseudo reinforcement learning algorithm has been introduced, which helps in reducing the skewness between the classes of the training dataset by appending the synthetic population of data across original minority class. The pseudo agent of the proposed algorithm efficiently fulfils an objective of minimisation of class skewness by sensing its overall state through immediate reward. By performing such task, the proposed algorithm satisfies the core idea of curtailing the training bias, which otherwise has negatively influenced the overall learning process. Further, the statistical analyses of the dataset suggest that, out of twelve features, the four clinical sub-features namely age, ejection fraction, serum creatinine and serum sodium along with follow-up time have significantly defined the relation with the death events. Hence, the aforementioned clinical sub-factors can be

utilised by the doctors to forecast the early signs of a possible survival of the patients due to LVSD.

ACKNOWLEDGEMENT

The authors would like to thank Prince Sultan University for their financial support.

REFERENCES

[1] Cardiovascular Diseases. (2017). *World Health Organization*. [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases

[2] T. Breen, B. Brueske, M. S. Sidhu, D. H. Murphree, K. B. Kashani, G. W. Barsness, and J. C. Jentzer, "Abnormal serum sodium is associated with increased mortality among unselected cardiac intensive care unit patients," *J. Amer. Heart Assoc.*, vol. 9, no. 2, Jan. 2020, Art. no. e014140.

[3] F. Formiga, R. Moreno-Gonzalez, D. Chivite, J. Franco, A. Montero, and X. Corbella, "High comorbidity, measured by the Charlson comorbidity index, associates with higher 1-year mortality risks in elderly patients experiencing a first acute heart failure hospitalization," *Aging Clin. Experim. Res.*, vol. 30, no. 8, pp. 927–933, Aug. 2018.

[4] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, p. 16, Dec. 2020.

[5] S. Choi, J. Chang, K. Kim, S. M. Kim, H.-Y. Koo, M. H. Cho, I. Y. Cho, H. Lee, J. S. Son, S. M. Park, and K. Lee, "Association of smoking cessation after atrial fibrillation diagnosis on the risk of cardiovascular disease: A cohort study of South Korean men," *BMC Public Health*, vol. 20, no. 1, p. 168, Dec. 2020.

[6] A. J. P. O. de Almeida, M. S. de Almeida Rezende, S. H. Dantas, S. de Lima Silva, J. C. P. L. de Oliveira, R. M. F. R. Alves, and G. M. S. de Menezes, "Unveiling the role of inflammation and oxidative stress on age-related cardiovascular diseases," *Oxid. Med. Cell. Longev.*, vol. 2020, pp. 1–20, May 2020.

[7] C. Franceschi, P. Garagnani, C. Morsiani, M. Conte, A. Santoro, A. Grignolio, D. Monti, M. Capri, and S. Salvioli, "The continuum of aging and age-related diseases: Common mechanisms but different rates," *Frontiers Med.*, vol. 5, Mar. 2018, Art. no. 61.

[8] Ada's Medical Knowledge Team. (2018). *Cardiovascular Disease Risk Factors*. [Online]. Available: http://www.ada.com/cardiovascular-disease-risk-factors

[9] D. E. Houghton, I. Koh, A. Ellis, N. S. Key, D. R. Douce, G. Howard, M. Cushman, M. Safford, and N. A. Zakai, "Hemoglobin levels and coronary heart disease risk by age, race, and sex in the reasons for geographic and racial differences in stroke study (REGARDS)," *Amer. J. Hematol.*, vol. 95, no. 3, pp. 258–266, Mar. 2020.

[10] A. Jayedi, S. Soltani, A. Abdolshahi, and S. Shab-Bidar, "Fish consumption and the risk of cardiovascular disease and mortality in patients with type 2 diabetes: A dose-response meta-analysis of prospective cohort studies," *Crit. Rev. Food Sci. Nutrition*, vol. 61, no. 10, pp. 1640–1650, May 2021, doi: 10.1080/10408398.2020.1764486.

[11] T. Kaisman-Elbaz, Y. Elbaz, V. Merkin, L. Dym, A. Noy, M. Atar-Vardi, R. Bari, S. Turiel, A. Alt, T. Zamed, Y. Eskira, K. Lavrenkov, Y. Kezerle, V. Dyomin, and I. Melamed, "Hemoglobin levels and red blood cells distribution width highlights glioblastoma patients subgroup with improved median overall survival," *Frontiers Oncol.*, vol. 10, Apr. 2020, Art. no. 432.

[12] W.-H. Lim, E.-K. Choi, K.-D. Han, S.-R. Lee, M.-J. Cha, and S. Oh, "Impact of hemoglobin levels and their dynamic changes on the risk of atrial fibrillation: A nationwide population-based study," *Sci. Rep.*, vol. 10, no. 1, p. 6762, Dec. 2020.

[13] S. Poludasu, K. Ramkissoon, L. Salciccioli, H. Kamran, and J. M. Lazar, "Left ventricular systolic function in sickle cell anemia: A meta-analysis," *J. Cardiac Failure*, vol. 19, no. 5, pp. 333–341, May 2013.

[14] M. Morissens, T. Besse-Hammer, M.-A. Azerad, A. Efira, and J. C. Rodriguez, "Evaluation of cardiac function in patients with sickle cell disease with left ventricular global longitudinal strain," *J. Transl. Internal Med.*, vol. 8, no. 1, pp. 41–47, May 2020.

[15] P. Angaran, "Association of left ventricular ejection fraction with mortality and hospitalizations," *J. Amer. Soc. Echocardiography*, vol. 33, pp. 802–811, Jul. 2020.

[16] B. Bagheri, N. Radmard, A. Makrani, and M. Rasouli, "Serum creatinine and occurrence and severity of coronary artery disease," *Med. Arch.*, vol. 73, no. 3, p. 154, 2019.

[17] A. Amin, M. Chitsazan, F. S. A. Abad, S. Taghavi, and N. Naderi, "On admission serum sodium and uric acid levels predict 30 day rehospitalization or death in patients with acute decompensated heart failure," *ESC Heart Failure*, vol. 4, no. 2, pp. 162–168, May 2017.

[18] A. F. Grubb, C. A. Pumill, S. J. Greene, A. Wu, K. Chiswell, and R. J. Mentz, "Tobacco smoking in patients with heart failure and coronary artery disease: A 20-year experience at duke university medical center," *Amer. Heart J.*, vol. 230, pp. 25–34, Dec. 2020.

[19] M. Metra, G. Cotter, M. Gheorghiade, L. D. Cas, and A. A. Voors, "The role of the kidney in heart failure," *Eur. Heart J.*, vol. 33, no. 17, pp. 2135–2142, Sep. 2012.

[20] S. Willoughby, "Platelets and cardiovascular disease," *Eur. J. Cardiovascular Nursing*, vol. 1, no. 4, pp. 273–288, Dec. 2002.

[21] Y.-J. Son and H.-J. Lee, "Association between persistent smoking after a diagnosis of heart failure and adverse health outcomes: A systematic review and meta-analysis," *Tobacco Induced Diseases*, vol. 18, Jan. 2020, Art. no. 5.

[22] S. J. Al'Aref, G. Singh, A. R. van Rosendael, K. K. Kolli, X. Ma, G. Maliakal, M. Pandey, B. C. Lee, J. Wang, Z. Xu, Y. Zhang, J. K. Min, S. C. Wong, and R. M. Minutello, "Determinants of in-hospital mortality after percutaneous coronary intervention: A machine learning approach," *J. Amer. Heart Assoc.*, vol. 8, no. 5, Mar. 2019, Art. no. e011160.

[23] E. G. Mansoori, "Using statistical measures for feature ranking," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 27, no. 1, Feb. 2013, Art. no. 1350003.

[24] J. Hoffmann, Y. Bar-Sinai, L. M. Lee, J. Andrejevic, S. Mishra, S. M. Rubinstein, and C. H. Rycroft, "Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets," *Sci. Adv.*, vol. 5, no. 4, Apr. 2019, Art. no. eaau6792.

[25] M. C. Jones, "Simple boundary correction for kernel density estimation," *Statist. Comput.*, vol. 3, no. 3, pp. 135–146, Sep. 1993.

[26] M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, J. van Langen, and R. A. Kievit, "Raincloud plots: A multi-platform tool for robust data visualization," *Wellcome Open Res.*, vol. 4, p. 63, Jan. 2021.

[27] E. O. Neftci and B. B. Averbeck, "Reinforcement learning in artificial and biological systems," *Nature Mach. Intell.*, vol. 1, no. 3, pp. 133–143, Mar. 2019.

[28] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, vol. 1, Cambridge, MA, USA: MIT Press, 1998.

[29] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 5th ed. Boca Raton, FL, USA: CRC Press, 2011, doi: 10.1201/9780429186196.

[30] T. Martinussen, S. Vansteelandt, M. Gerster, and J. V. B. Hjelmborg, "Estimation of direct effects for survival data by using the aalen additive hazards model," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 73, no. 5, pp. 773–788, Nov. 2011.

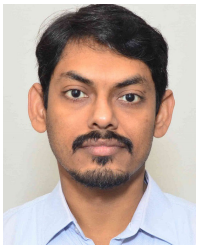
[31] D. W. Hosmer and P. Royston, "Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model," *Stata J., Promoting Commun. Statist. Stata*, vol. 2, no. 4, pp. 331–350, Dec. 2002.

[32] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics Proteomics*, vol. 15, pp. 41–51, Jan. 2018.

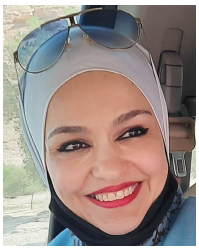
MOHAMMAD FARHAN KHAN received the B.Tech. and M.Tech. degrees in electronics engineering from the Z. H. College of Engineering and Technology, Aligarh Muslim University, India, in 2010 and 2012, respectively; and the Ph.D. degree in electronic engineering from the School of Engineering and Digital Arts, University of Kent, U.K., in 2017. He has worked as a Postdoctoral Research Associate in the EPSRC project that was in joint collaborated between the University of Warwick, U.K., and the University of Central Lancashire, U.K., and also worked as a MHRD Fellow with the Indian Institute of Technology Roorkee, India. He is currently working as an Innovate U.K. Research Fellow with Cranfield University, U.K. His research interests include control theory application, 3D monitoring, robotic vision systems, image processing, mathematical modeling, soft computing, applied machine learning, and computational biology.

RAJESH KUMAR GAZARA received the B.Sc. degree in bioinformatics from Jaipur National University, in 2011, the M.Sc. degree in bioinformatics from Jamia Millia Islamia, India, in 2013, and the Ph.D. degree in bioscience and biotechnology from the Universidade Estadual do Norte Fluminense Darcy Ribeiro, Brazil, in 2019. He has worked as a Postdoctoral Fellow with the Indian Institute of Technology Roorkee, India. His research interests include genomics, comparative genomics, transcriptomics, next-generation sequencing, sequence analysis, computational biology, and statistical analysis.

MUAFFAQ M. NOFAL received the B.S. and M.S. degrees in physics from the University of Jordan, in 1991 and 1995, respectively, and the Ph.D. degree in experimental atomic physics from Frankfurt University, Germany, in 2007. From 2007 to 2009, he worked as an Assistant Professor with Applied Science University, Amman, Jordan. Since 2009, he has been an Assistant Professor with the Science Department, Prince Sultan University, Saudi Arabia.



SOHOM CHAKRABARTY received the B.E. degree in electrical engineering from Jadavpur University, India, in 2008, and the Ph.D. degree in control systems from IIT Bombay, India, in 2015. He is currently working as an Assistant Professor with IIT Roorkee. His research interests include sliding mode control, learning-based bioinformatics, and multi-agent systems.



ELHAM M. A. DANNOUN received the M.Sc. and Ph.D. degrees in physics from the University of Jordan, Amman, Jordan. She worked as an Assistant Professor with HCT, Dubai, United Arab Emirates. She is currently working as the Associate Director of the General Science Department, Prince Sultan University, Riyadh, Saudi Arabia. Her research interests include mathematical physics, material science, magnetism, and electrical engineering.

RAMI AL-HMOUZ received the B.Sc. degree in electrical engineering from Mutah University, Alkarak, Jordon, in 1998, the M.Sc. degree in computer engineering from the University of Western Sydney, Sydney, NSW, Australia, in 2004, and the Ph.D. degree in computer engineering from the University of Technology, Sydney, NSW, in 2008. He is currently working as an Associate Professor with Sultan Qaboos University, Oman. His research interests include machine learning, computer vision, and granular computing.

M. MURSALEEN has worked as a Chairman of the Department of Mathematics, Aligarh Muslim University, where he is currently a Principal Investigator for a SERB Core Research Project. He is also affiliated with China Medical University, Taichung, Taiwan. He has published nine books and more than 330 research articles in the field of summability, sequence spaces, approximation theory, fixed point theory, and measures of noncompactness. He has a number of academic visits in several countries and has successfully completed several national and international projects. Besides mentoring several master's students, he has guided twenty Ph.D. students. He is acting as a member of editorial boards for several international scientific journals. He also served as a reviewer for various international scientific journals. He has been recognised as the Highly Cited Researcher (for year 2019) by the Web of Science.